



DH06

Transparent is the New Black

Improving Data Findability
Through Better Clinical Metadata

Łukasz Knioła
Biogen

PHUSE EU Connect 2023, Birmingham

Disclaimer

Any views, thoughts, and opinions stated in this presentation are those of the author alone, and they **do not represent the positions or describe processes** of their employer or any other organization or group they may be affiliated with.



“F” in F.A.I.R. stands for FINDABLE



“F” in F.A.I.R. stands for FINDABLE

~~FIND~~
DISCOVER



FOUND

ALPHA



GAMMA



SIGMA



HbA1c



FOUND

ALPHA



GAMMA



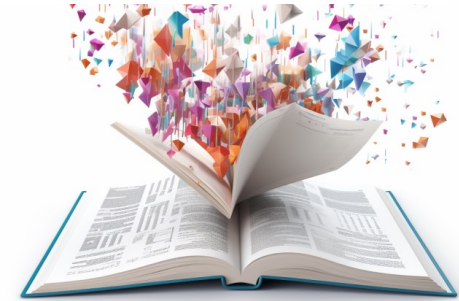
SIGMA



HbA1c

UNDISCOVERED

KAPPA



LAMBDA



DISCOVER the unknown



Different flavours of metadata

```
<ItemGroupDef OID="IG.VS" Domain="VS" Name="VS" Repeating="Yes" IsReferenceData="No"
  SASDatasetName="VS"
  Purpose="Tabulation"
  def:Structure="One record per vital sign measurement per time point per visit per subject"
  def:Class="FINDINGS"
  def:ArchiveLocationID="LF.VS">
  <ItemRef ItemOID="IT.VS.VSTESTCD" OrderNumber="5" Mandatory="N"
</ItemGroupDef>
```

```
<ModelList OID="CL.VSTESTCD" Name="VSTESTCD" DataType="text">
  <Model CodedValue="HEIGHT" OrderNumber="3">
    <TranslatedText>Height</TranslatedText>
```

```
<ItemDef OID="IT.VS.VSTESTCD" Name="VSTESTCD" DataType="text">
  SASFieldName="VSTESTCD">
  <Description>
    <TranslatedText xml:lang="en">Vital Sign Test
```

Library Name	Member Name	Member Type	Column Name	Column Type	Column Length	Column Position	Column Number in Table	Column Label
DATA	DM	DATA	STUDYID	char	8	8	1	Study Identifier
DATA	DM	DATA	DOMAIN	char	2	16	2	Domain Abbreviation
DATA	DM	DATA	USUBJID	char	19	18	3	Unique Subject Identifier
DATA	DM	DATA	SUBJID	char	10	37	4	Subject Identifier for the Study
DATA	DM	DATA	RFSTDTC	char	19	47	5	Subject Reference Start Date/Time
DATA	DM	DATA	RFENDTC	char	19	66	6	Subject Reference End Date/Time
DATA	DM	DATA	SITEID	char	10	85	7	Study Site Identifier
DATA	DM	DATA	BRTHDTC	char	19	95	8	Date/Time of Birth
DATA	DM	DATA	AGE	num	8	0	9	Age in AGEU at RFSTDTC
DATA	DM	DATA	AGEU	char	6	114	10	Age Units
DATA	DM	DATA	SEX	char	8	120	11	Sex
DATA	DM	DATA	RACE	char	47	128	12	Race
DATA	DM	DATA	ARMCD	char	20	175	13	Planned Arm Code
DATA	DM	DATA	ARM	char	30	195	14	Description of Planned Arm
DATA	DM	DATA	COUNTRY	char	3	225	15	Country

```
-bash-4.2$ ls -l --full-time
total 5256
-rwxr-xr-x 1 lkniola lkniola 176597 2018-09-10 15:10:37.000000000 -0400 ae.csv
-rwxr-xr-x 1 lkniola lkniola 958464 2019-12-02 11:08:18.175240500 -0500 ae.sas7bdat
-rwxr-xr-x 1 lkniola lkniola 16227 2020-07-01 12:13:118500 -0400 dm.csv
-rwxr-xr-x 1 lkniola lkniola 13107 2020-07-01 12:000000000 -0500 dm.sas7bdat
-rwxr-xr-x 1 lkniola lkniola 733648300 -0400 ex.csv
-rwxr-xr-x 1 lkniola lkniola 000000000 -0500 ex.sas7bdat
-rwxr-xr-x 1 lkniola lkniola 454165018 -0400 vs.csv
-rwxr-xr-x 1 lkniola lkniola 454165018 -0400 vs.sas7bdat
```

Different flavours of metadata

Study Level Metadata (ClinicalTrial.gov, ...)



Result Level Metadata

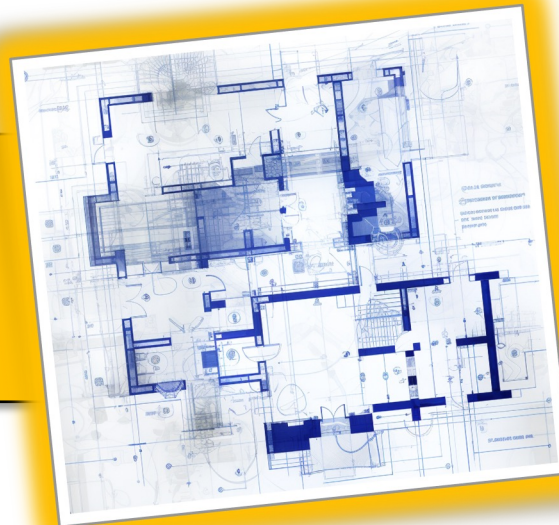


Dataset Level Metadata

SDTM mapping

ADaM derivation

CT + VLM



Can we create, extract or otherwise make available
Result Level Metadata for any study and dataset we have?

How can we do it in an automated way instead of processing each study and dataset separately (manually)?

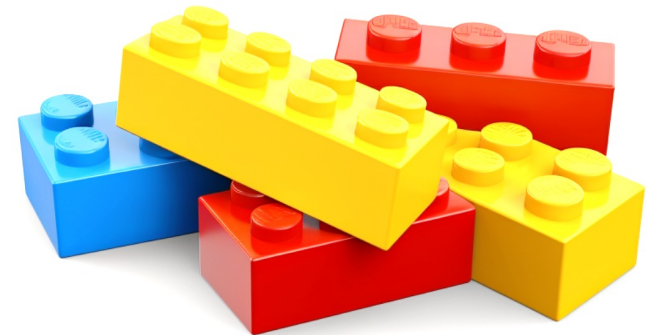
The stackability of SDTM datasets

Structure:

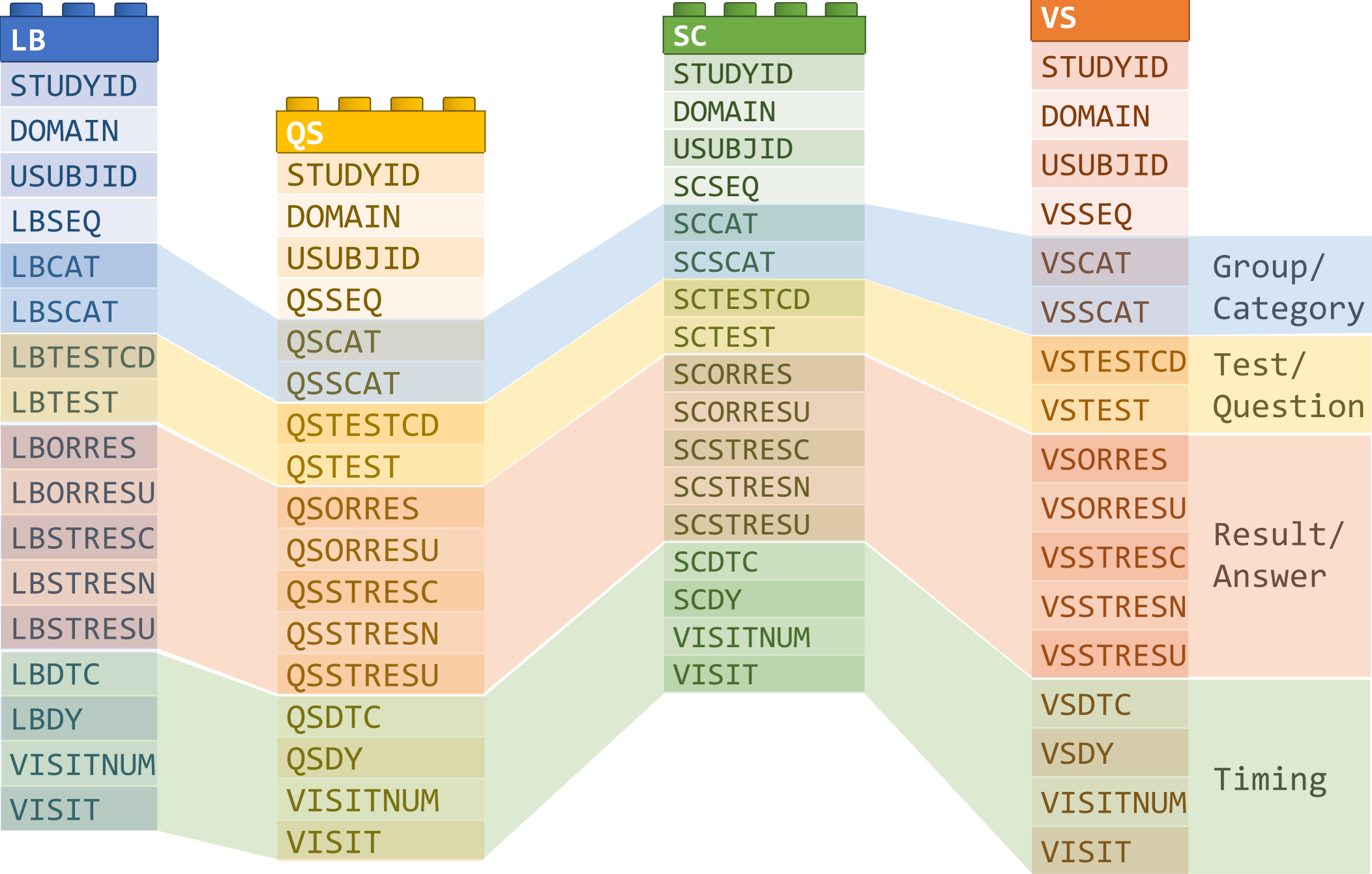
- IDs (study, subject, ...)
- Category / internal grouping of records
- Question, test, assessment
- Answer, result, finding, observation
- Timing in the form of dates/study days
- Timing in the form of study period, stage, visit
- Domain specific details

SDTM domain models:

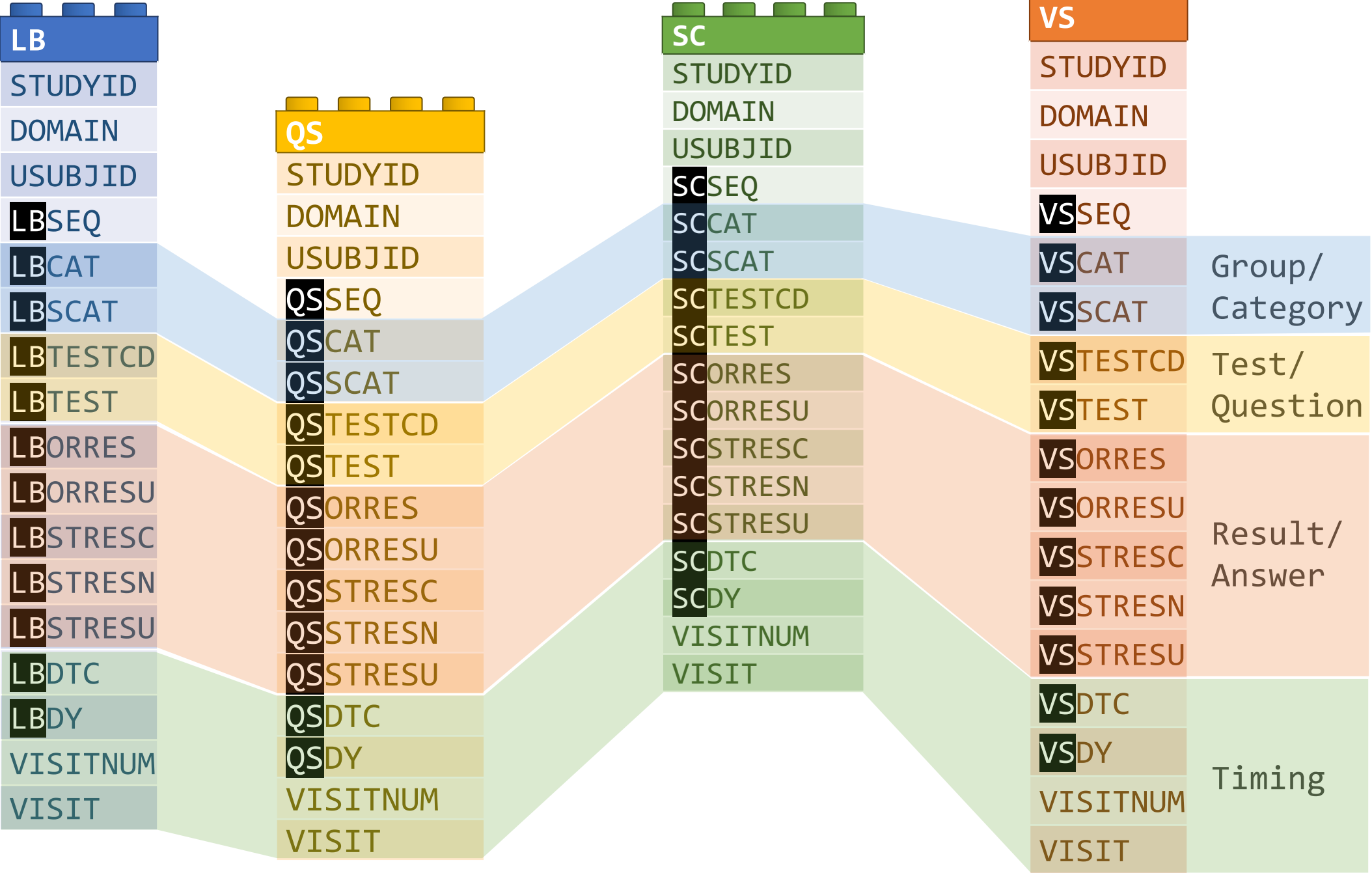
- Special purpose domains (DM, CO, SE, SV)
- Interventions (CM, EX, SU)
- Events (AE, DS, MH, DV, CE)
- Findings (EG, IE, LB, PE, QS, SC, VS, etc.)
- Trial design domains (TA, TE, TV, TI, TS)



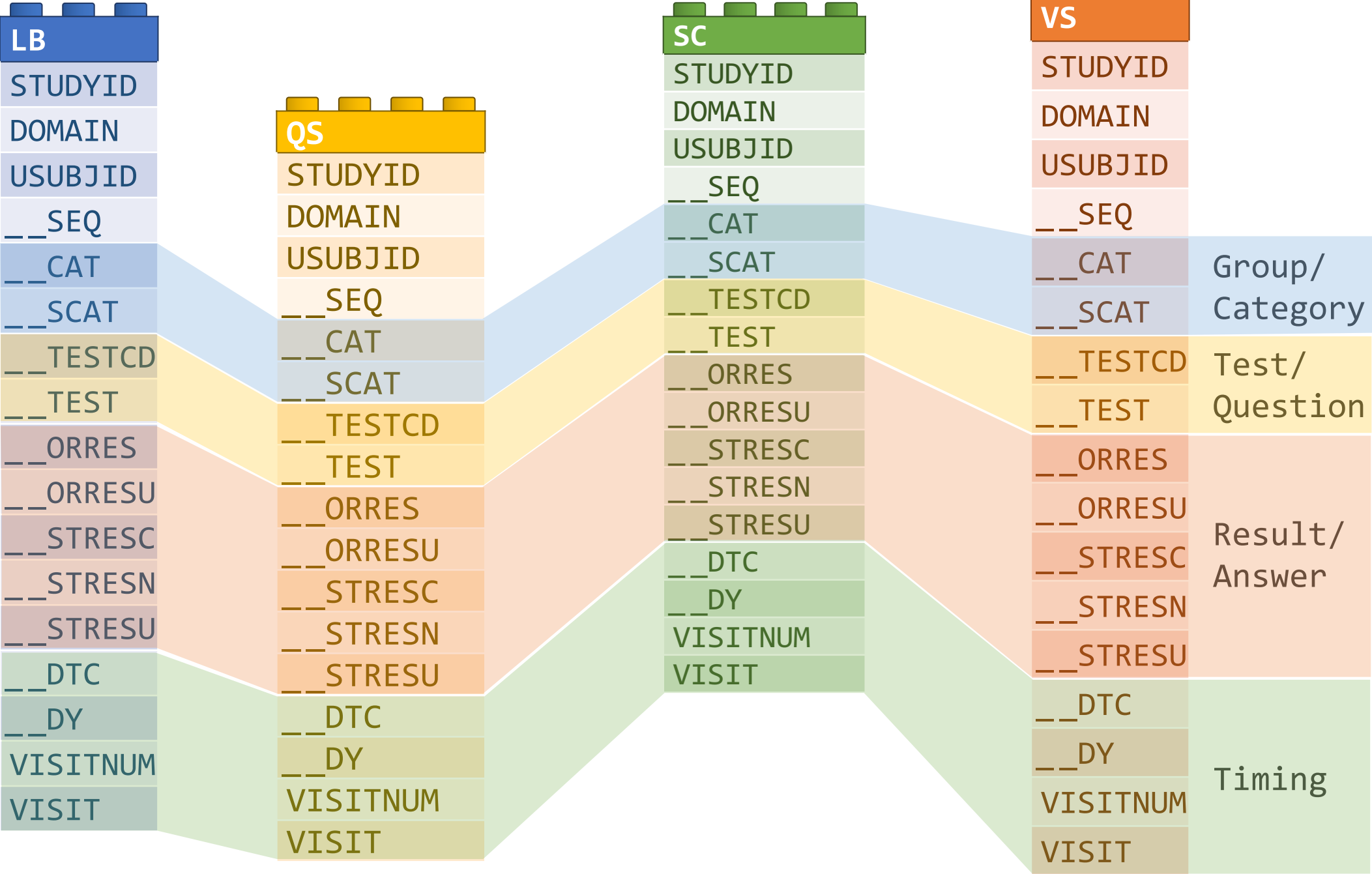
Findings



Findings



Findings



Findings

LB	QS	SC	VS		
STUDYID	STUDYID	STUDYID	STUDYID		
DOMAIN	DOMAIN	DOMAIN	DOMAIN		
USUBJID	USUBJID	USUBJID	USUBJID		
__SEQ	__SEQ	__SEQ	__SEQ		
__CAT	__CAT	__CAT	__CAT	Group/ Category	
__SCAT	__SCAT	__SCAT	__SCAT		
__TESTCD	__TESTCD	__TESTCD	__TESTCD	Test/ Question	
__TEST	__TEST	__TEST	__TEST		
__ORRES	__ORRES	__ORRES	__ORRES	Result/ Answer	
__ORRESU	__ORRESU	__ORRESU	__ORRESU		
__STRESC	__STRESC	__STRESC	__STRESC		
__STRESN	__STRESN	__STRESN	__STRESN		
__STRESU	__STRESU	__STRESU	__STRESU		
__DTC	__DTC	__DTC	__DTC	Timing	
__DY	__DY	__DY	__DY		
VISITNUM	VISITNUM	VISITNUM	VISITNUM		
VISIT	VISIT	VISIT	VISIT		

Findings

FINDINGS		
STUDYID		
DOMAIN		
USUBJID		
__SEQ		
__CAT	Group/ Category	
__SCAT		
__TESTCD	Test/ Question	
__TEST		
__ORRES	Result/ Answer	
__ORRESU		
__STRESC		
__STRESN		
__STRESU		
__DTC	Timing	
__DY		
VISITNUM		
VISIT		

Findings

FINDINGS		
STUDYID		
DOMAIN		
USUBJID		
__CAT		Group/ Category
__SCAT		
__TESTCD		Test/ Question
__TEST		
__STRESC		Result/ Answer
__STRESN		
__STRESU		
__DTC		Timing
__DY		
VISIT		

Findings + Interventions + Events

FINDINGS		INTERVENTIONS		EVENTS	
STUDYID		STUDYID		STUDYID	
DOMAIN		DOMAIN		DOMAIN	
USUBJID		USUBJID		USUBJID	
__CAT		__CAT		__CAT	Group/ Category
__SCAT		__SCAT		__SCAT	
__TESTCD		DOMAIN		DOMAIN	Test/ Question
__TEST		(dataset label)		(dataset label)	
__STRESC		__TRT		__TRT	Result/ Answer
__STRESN		__DECOD		__DECOD	
__STRESU		__OCCUR		__BODSYS	
__DTC		__STDTC, __ENDTC		__STDTC, __ENDTC	Timing
__DY		__STDY, __ENDY		__STDY, __ENDY	
VISIT		__ENRF		__ENRF	

Demog

DM (AGE)		DM (SEX)		
STUDYID		STUDYID		
DOMAIN		DOMAIN		
USUBJID		USUBJID		
				Group/Category
"AGEGRP"		"SEX"		Test/Question
grp(AGE)*		SEX		Result/Answer
"BASELINE"		"BASELINE"		Timing

↑
Repeat for COUNTRY, RACE, ARM, etc.

*grp(AGE): ceil(AGE/10)*10 || "-" || (ceil(AGE/10)*10)-9

Extracting Result-Level Metadata

```
create table RLM_FINDINGS as
  select distinct STUDYID, DOMAIN
                , __CAT, __SCAT
                , __TESTCD, __TEST

                , count(*) as N
                , count(distinct USUBJID) as N_SUBJ

                , mean(__STRESN) as MEAN, median(__STRESN) as MEDIAN
                , min(__STRESN) as MIN, max(__STRESN) as MAX
  from FINDINGS

  group by STUDYID, DOMAIN, __CAT, __SCAT, __TESTCD, __TEST
;
```

VARs:

ID

GROUP

QUESTION

ANSWER

TIMING

Extracting Result-Level Metadata

VARs:

ID

GROUP

QUESTION

ANSWER

TIMING

```
create table RLM_FINDINGS as
  select distinct STUDYID, DOMAIN
    , catx(' :: ', __CAT, __SCAT ) as GRP
    , catx(' :: ', __TESTCD, __TEST ) as Q
    , __STRESU as UNIT
    , VISIT as VISIT
    , count(*) as N
    , count(distinct USUBJID) as N_SUBJ

    , mean(__STRESN) as MEAN, median(__STRESN) as MEDIAN
    , min(__STRESN) as MIN, max(__STRESN) as MAX
  from FINDINGS

  where __STRESC <> ''
  group by STUDYID, DOMAIN, __CAT, __SCAT, __TESTCD, __TEST, __STRESU, VISIT
;
```

Extracting Result-Level Metadata

```
create table RLM_FINDINGS as
  select distinct STUDYID, DOMAIN
    , catx(' :: ', __CAT, __SCAT ) as GRP
    , catx(' :: ', __TESTCD, __TEST ) as Q
    , __STRESU as UNIT
    , VISIT as VISIT
    , count(*) as N
    , count(distinct USUBJID) as N_SUBJ
    , (calculated N_SUBJ)/_TOT_SUBJ as P
    , mean(__STRESN) as MEAN, median(__STRESN) as MEDIAN
    , min(__STRESN) as MIN, max(__STRESN) as MAX
  from FINDINGS
    , (select count(distinct USUBJID) as _TOT_SUBJ from DM)
  where __STRESC <> ''
  group by STUDYID, DOMAIN, __CAT, __SCAT, __TESTCD, __TEST, __STRESU, VISIT
;
```

VARs:

ID

GROUP

QUESTION

ANSWER

TIMING

Output Result-Level Metadata

FINDINGS

STUDYID	DOMAIN	GRP	Q	UNIT	N	N_SUBJ	P	MEAN	MEDIAN	MIN	MAX
PHU23	VS	BLOOD PRESSURE	DIABP :: Diastolic Blood Pressure	mmHg	2071	150	1.000	99.3	99	61	140
PHU23	VS	BLOOD PRESSURE	PULSE :: Pulse Rate	bpm	2071	150	1.000	111.2	111	48	173
PHU23	VS	BLOOD PRESSURE	SYSBP :: Systolic Blood Pressure	mmHg	2071	150	1.000	120.5	120	71	170
PHU23	VS	BODY MEASUREMENT	HEIGHT :: Height	cm	148	148	0.986	175.1	175	152	201
PHU23	VS	BODY MEASUREMENT	WEIGHT :: Weight	kg	2043	148	0.986	88.3	88	47	131

Output Result-Level Metadata

EVENTS

STUDYID	DOMAIN	GRP	Q	UNIT	N	N_SUBJ	P	MEAN	MEDIAN	MIN	MAX
PHU23	AE	ADVERSE EVENT	BACK PAIN		39	34	0.226				
PHU23	AE	ADVERSE EVENT	DIZZINESS		9	9	0.060				
PHU23	AE	ADVERSE EVENT	HEADACHE		138	87	0.580				
PHU23	AE	ADVERSE EVENT	INJECTION SITE PAIN		14	14	0.093				
PHU23	AE	ADVERSE EVENT	NASOPHARYNGITIS		59	50	0.333				
PHU23	AE	ADVERSE EVENT	PARAESTHESIA		17	16	0.106				
PHU23	AE	ADVERSE EVENT	URINARY TRACT INFECTION		38	32	0.213				

Output Result-Level Metadata

DEMOG

STUDYID	DOMAIN	GRP	Q	UNIT	N	N_SUBJ	P	MEAN	MEDIAN	MIN	MAX
PHU23	DM	AGEGRP	21-30		112	112	0.746				
PHU23	DM	AGEGRP	31-40		38	38	0.253				
PHU23	DM	COUNTRY	GBR		55	55	0.366				
PHU23	DM	COUNTRY	IND		50	50	0.333				
PHU23	DM	COUNTRY	USA		45	45	0.300				
PHU23	DM	SEX	F		81	81	0.540				
PHU23	DM	SEX	M		69	69	0.460				
PHU23	DM	TOTCOUNT	TOTAL COUNT		150	150	1.000				
		EVENT	INFECTION								

Stacking Result-Level Metadata

STUDYID	DOMAIN	GRP	Q	UNIT	N	N_SUBJ	P	MEAN	MEDIAN	MIN	MAX
---------	--------	-----	---	------	---	--------	---	------	--------	-----	-----

PHU23	AE	ADVERSE EVENT	BACK PAIN		39	34	0.226				
PHU23	AE	ADVERSE EVENT	DIZZINESS		9	9	0.060				
PHU23	AE	ADVERSE EVENT	HEADACHE		138	87	0.580				
PHU23	AE	ADVERSE EVENT	INJECTION SITE PAIN		14	14	0.093				
PHU23	AE	ADVERSE EVENT	NASOPHARYNGITIS		59	50	0.333				
PHU23	AE	ADVERSE EVENT	PARAESTHESIA		17	16	0.106				
PHU23	AE	ADVERSE EVENT	URINARY TRACT INFECTION		38	32	0.213				

EVENTS

PHU23	DM	AGEGRP	21-30		112	112	0.746				
PHU23	DM	AGEGRP	31-40		38	38	0.253				
PHU23	DM	COUNTRY	GBR		55	55	0.366				
PHU23	DM	COUNTRY	IND		50	50	0.333				
PHU23	DM	COUNTRY	USA		45	45	0.300				
PHU23	DM	SEX	F		81	81	0.540				
PHU23	DM	SEX	M		69	69	0.460				
PHU23	DM	TOTALCOUNT	TOTAL COUNT		150	150	1.000				

DEMOG

PHU23	VS	BLOOD PRESSURE	DIABP :: Diastolic Blood Pressure	mmHg	2071	150	1.000	99.3	99	61	140
PHU23	VS	BLOOD PRESSURE	PULSE :: Pulse Rate	bpm	2071	150	1.000	111.2	111	48	173
PHU23	VS	BLOOD PRESSURE	SYSBP :: Systolic Blood Pressure	mmHg	2071	150	1.000	120.5	120	71	170
PHU23	VS	BODY MEASUREMENT	HEIGHT :: Height	cm	148	148	0.986	175.1	175	152	201
PHU23	VS	BODY MEASUREMENT	WEIGHT :: Weight	kg	2043	148	0.986	88.3	88	47	131

FINDINGS

Stacking Result-Level Metadata

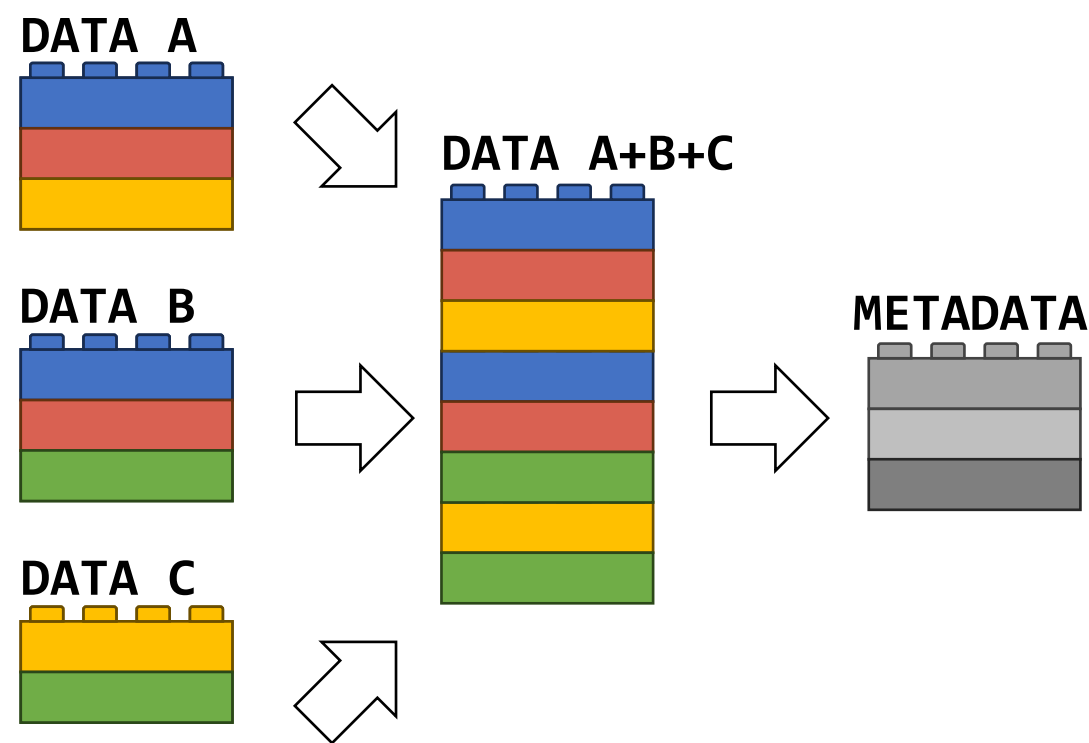
STUDYID	DOMAIN	GRP	Q	UNIT	N	N_SUBJ	P	MEAN	MEDIAN	MIN	MAX
PHU23	AE	ADVERSE EVENT	BACK PAIN		39	34	0.226				
PHU23	AE	ADVERSE EVENT	DIIZZINESS		9	9	0.060				
PHU23	AE	ADVERSE EVENT	HEADACHE		138	87	0.580				
PHU23	AE	ADVERSE EVENT	INJECTION SITE PAIN		14	14	0.093				
PHU23	AE	ADVERSE EVENT	NASOPHARYNGITIS		59	50	0.333				
PHU23	AE	ADVERSE EVENT	PARAESTHESIA		17	16	0.106				
PHU23	AE	ADVERSE EVENT	URINARY TRACT INFECTION		38	32	0.213				
PHU23	DM	AGEGRP	21-30		112	112	0.746				
PHU23	DM	AGEGRP	31-40		38	38	0.253				
PHU23	DM	COUNTRY	GBR		55	55	0.366				
PHU23	DM	COUNTRY	IND		50	50	0.333				
PHU23	DM	COUNTRY	USA		45	45	0.300				
PHU23	DM	SEX	F		81	81	0.540				
PHU23	DM	SEX	M		69	69	0.460				
PHU23	DM	TOTALCOUNT	TOTAL COUNT		150	150	1.000				
PHU23	VS	BLOOD PRESSURE	DIABP :: Diastolic Blood Pressure	mmHg	2071	150	1.000	99.3	99	61	140
PHU23	VS	BLOOD PRESSURE	PULSE :: Pulse Rate	bpm	2071	150	1.000	111.2	111	48	173
PHU23	VS	BLOOD PRESSURE	SYSBP :: Systolic Blood Pressure	mmHg	2071	150	1.000	120.5	120	71	170
PHU23	VS	BODY MEASUREMENT	HEIGHT :: Height	cm	148	148	0.986	175.1	175	152	201
PHU23	VS	BODY MEASUREMENT	WEIGHT :: Weight	kg	2043	148	0.986	88.3	88	47	131
PHU23	LB										
PHU23	LB										
PHU23	LB										
PHU23	LB										
PHU23	LB										
PHU23	LB										
PHU23	LB										
PHU23	CM										
PHU23	CM										
PHU23	CM										
PHU23	CM										
PHU23	CM										
PHU23	CM										
PHU24	DM										
PHU24	DM										
PHU24	DM										
PHU24	DM										
PHU24	DM										
PHU24	DM										

EVENTS

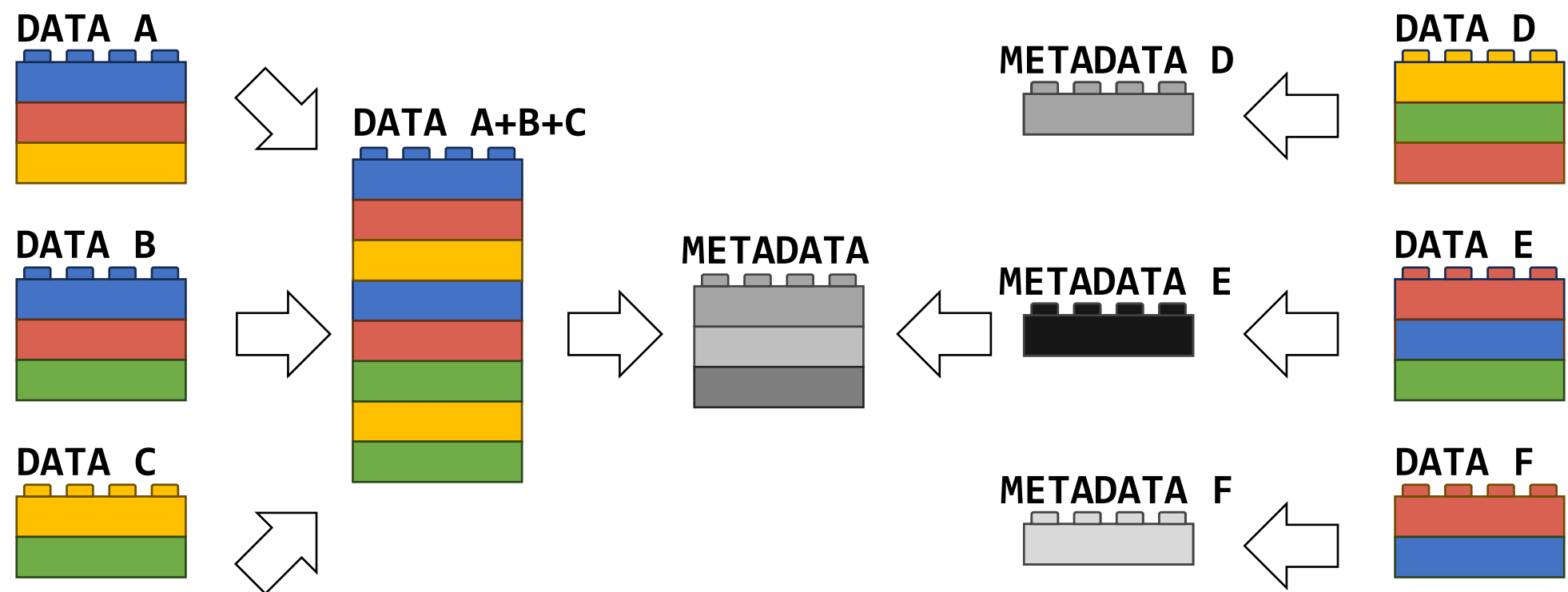
DEMOG

FINDINGS

Processing stacked datasets

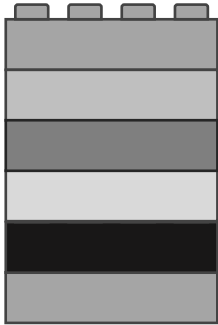


Processing stacked datasets



Processing stacked datasets

METADATA



```
show : STUDYID, N_SUBJ  
where: Q=HbA1c
```

```
result:
```

```
STUDYID | N_SUBJ
```

```
-----+-----
```

```
ALPHA   |    245
```

```
GAMMA   |    187
```

```
KAPPA   |    309
```

```
LAMBDA  |    348
```

```
SIGMA   |    192
```

```
> _
```



```
show : STUDYID, Q, MEAN, MIN, MAX, UNIT
where: BASEPOST=baseline & Q=HbA1c
```

result:

STUDYID	Q	MEAN	MIN	MAX	UNIT
ALPHA	HbA1c	38.3	34	47	mmol/mol
GAMMA	HbA1c	40.1	36	51	mmol/mol
KAPPA	HbA1c	43.5	40	49	mmol/mol
LAMBDA	HbA1c	39.9	35	42	mmol/mol
SIGMA	HbA1c	42.0	33	46	mmol/mol

```
> _
```



```
show : STUDYID, GRP, Q, N, P
where: STUDY in (KAPPA, LAMBDA) & DOMAIN=DM
```

result:

STUDYID	GRP	Q	N	P
KAPPA	AGEGRP	21-30	142	0.4595
KAPPA	AGEGRP	31-40	167	0.5404
KAPPA	COUNTRY	GBR	122	0.3948
KAPPA	COUNTRY	IND	187	0.6051
KAPPA	RACE	ASIAN	201	0.6504
KAPPA	RACE	WHITE	108	0.3495
KAPPA	SEX	F	162	0.5242
KAPPA	SEX	M	147	0.4757
KAPPA	TOTCNT	TOTAL	309	1.0000
LAMBDA	AGEGRP	21-30	155	0.5016
LAMBDA	AGEGRP	31-40	193	0.6245
LAMBDA	COUNTRY	GBR	131	0.4230



Use the metadata

Which studies had participants in France and/or Spain?

```
>STUDYID, N_SUBJ: GRP=COUNTRY & (Q=FRA | Q=ESP) & N>0
```

Which studies had participants between 41 and 50 years old?

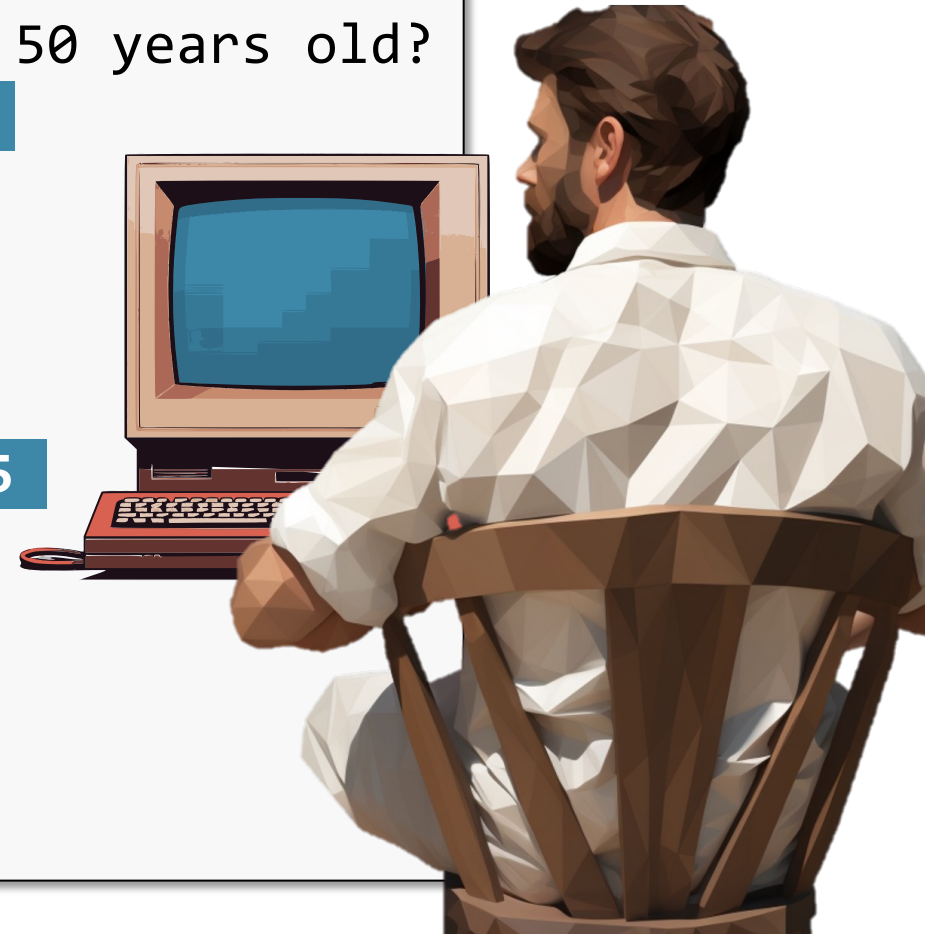
```
>STUDYID, N_SUBJ: GRP=AGEGRP & Q=41-50 & N>0
```

Have there been studies where prevalence of cough was >50% among participants?

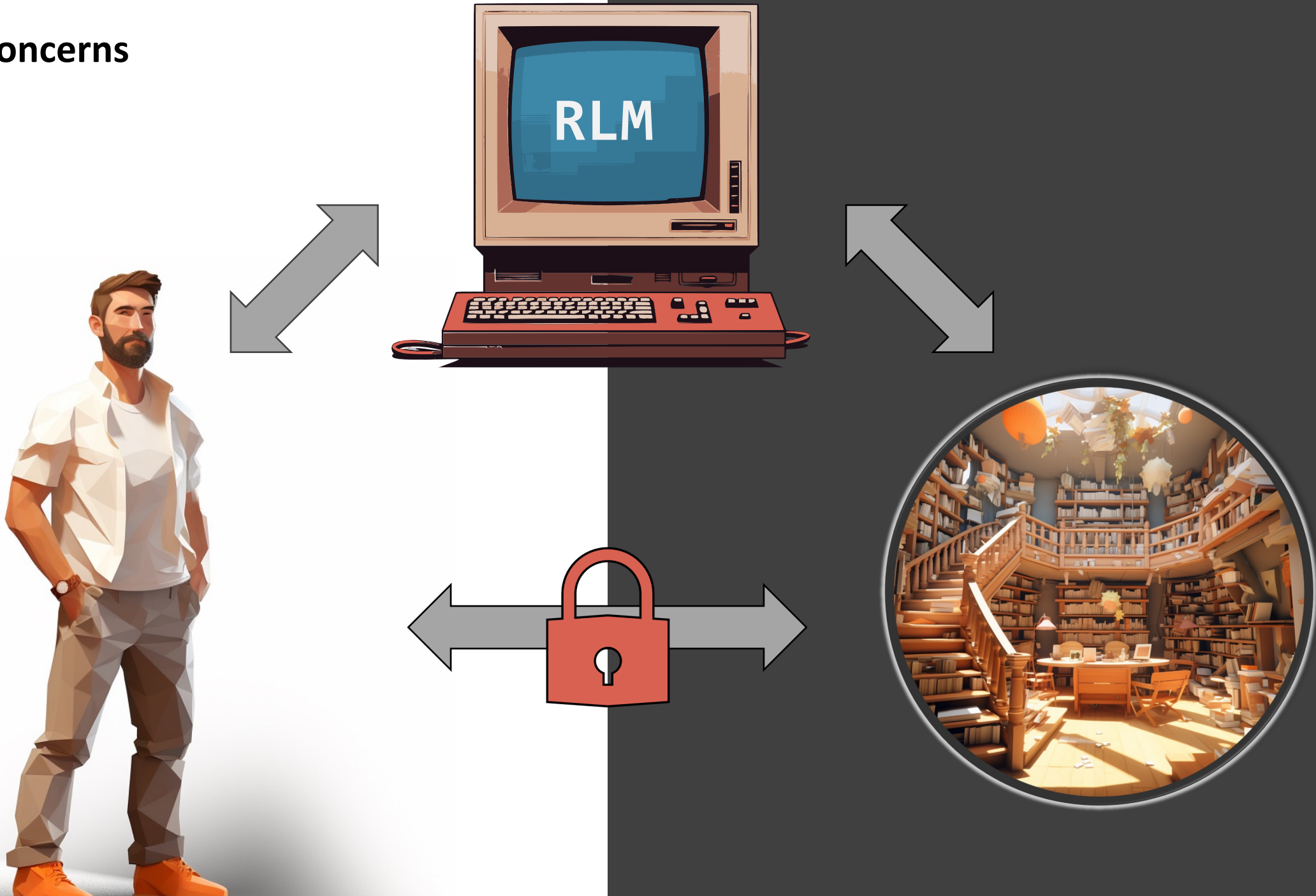
```
>STUDYID, N_SUBJ: DOMAIN=AE & Q=COUGH & P>0.5
```

Which datasets contain MRI results?

```
>DOMAIN: GRP like 'MRI%' | Q like 'MRI%'
```



Privacy concerns

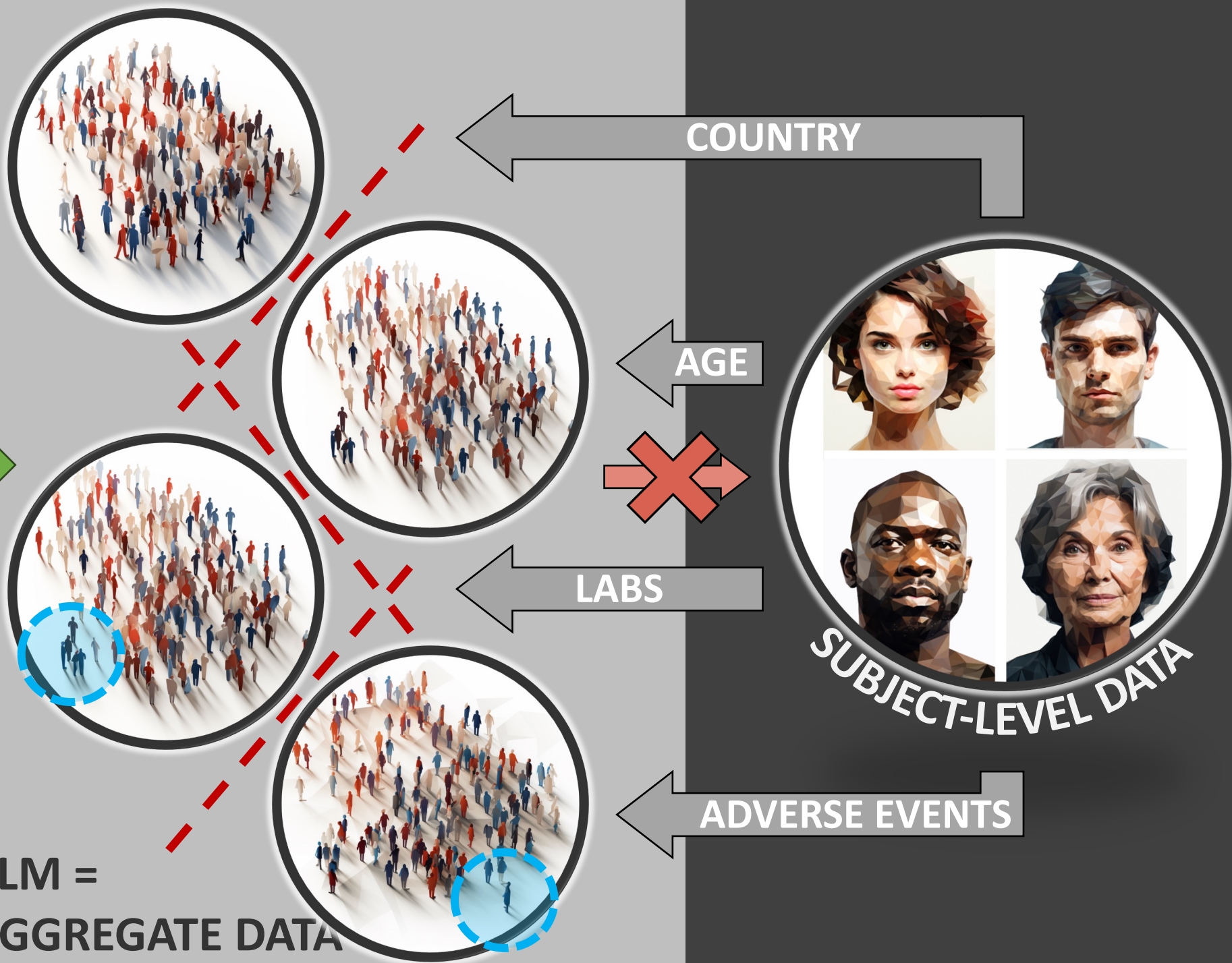


Privacy concerns

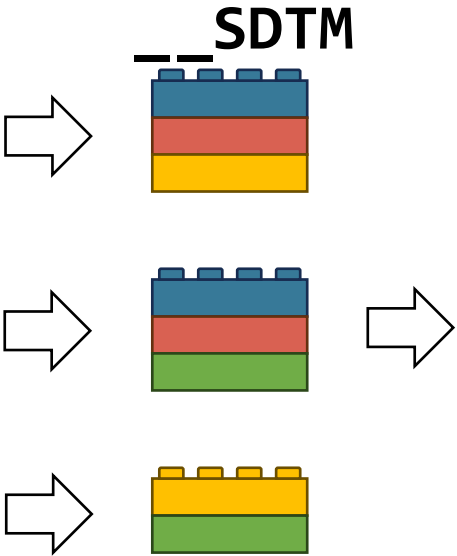
- ✓ Linkability
- ✓ Singling out
- ✓ Inference



RLM =
AGGREGATE DATA



Result Level Metadata - Summary



RLM

STUDYID	DOMAIN	GRP	Q	UNIT	N	N_SUBJ	P	MEAN	MEDIAN	MIN	MAX
PHL_023	AE	ADVERSE EVENT	BACK PAIN		59	34	0.2267				
PHL_023	AE	ADVERSE EVENT	DIZZINESS		9	9	0.0608				
PHL_023	AE	ADVERSE EVENT	HEADACHE		138	87	0.5888				
PHL_023	AE	ADVERSE EVENT	INJECTION SITE PAIN		14	14	0.0933				
PHL_023	AE	ADVERSE EVENT	NASOPHARYNGITIS		59	58	0.3333				
PHL_023	AE	ADVERSE EVENT	PARAESTHESIA		17	16	0.1067				
PHL_023	AE	ADVERSE EVENT	URINARY TRACT INFECTION		38	32	0.2133				
PHL_023	DM	AGEGRP	21-39		112	112	0.7467				
PHL_023	DM	AGEGRP	31-49		38	38	0.2533				
PHL_023	DM	COUNTRY	GBR		55	55	0.3667				
PHL_023	DM	COUNTRY	IND		58	58	0.3333				
PHL_023	DM	COUNTRY	USA		45	45	0.3008				
PHL_023	DM	RACE	ASIAN		43	43	0.2867				
PHL_023	DM	RACE	BLACK OR AFRICAN AMERICAN		14	14	0.0933				
PHL_023	DM	RACE	WHITE		93	93	0.6208				
PHL_023	DM	SEX	F		61	61	0.5408				
PHL_023	DM	SEX	M		69	69	0.4592				
PHL_023	DM	TOTALCOUNT	TOTAL COUNT		158	158	1.0000				
PHL_023	VS	BLOOD PRESSURE	DIABP :: Diastolic Blood Pressure	mmHg	2671	158	1.0008	99.385	99	61	148
PHL_023	VS	BLOOD PRESSURE	PULSE :: Pulse Rate	beats/min	2671	158	1.0008	111.215	111	48	173
PHL_023	VS	BLOOD PRESSURE	SYSBP :: Systolic Blood Pressure	mmHg	2671	158	1.0008	120.524	128	71	178
PHL_023	VS	BODY MEASUREMENT	HEIGHT :: Height	cm	148	148	0.9867	175.142	175	152	201
PHL_023	VS	BODY MEASUREMENT	WEIGHT :: Weight	kg	2643	148	0.9867	88.354	88	47	131



THANK YOU!

DH06

Transparent is the New Black

Improving Data Findability
Through Better Clinical
Metadata

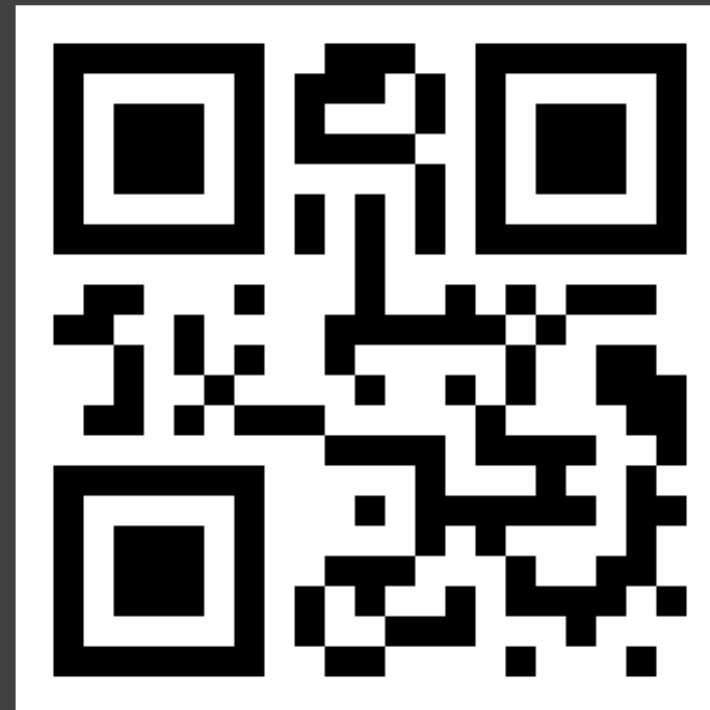
Łukasz Knioła

Biogen

✉ lukasz.kniola@biogen.com

PHUSE EU Connect 2023

CODE:



t.ly/tZW8P